Short Text Classification in Twitter through a semantic transform based on Wikipedia

by
SOMESH JAIN

Under the Supervision of

PROF. NILOY GANGULY

Submitted in fulfilment of the requirements for award of the degree of

Master of Technology
in
Computer Science and Engineering

Department of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Declaration by the Author of the Thesis

I, **Somesh Jain**, Roll No **08CS3013**, registered as a student of 5 year Dual Degree Program (B.Tech. + M.Tech.) in the Department of **Computer Science And Engineering**, Indian Institute of Technology, Kharagpur, India (hereinafter referred to as the 'Institute') do hereby submit my thesis, titled:

Short Text Classification in Twitter through a Semantic Transform based on Wikipedia Applications (hereinafter referred to as 'my thesis') in a printed as well as in an electronic version for holding in the library of record of the Institute.

I hereby declare that:

1. The electronic version of my thesis submitted herewith on CDROM is in PDF Format.

2. My thesis is my original work of which the copyright vests in me and my thesis do not infringe or violate the rights of anyone else.

3. The contents of the electronic version of my thesis submitted herewith are the same as that submitted as final hard copy of my thesis after my viva voce and adjudication of my thesis in **April 2013**.

4. I agree to abide by the terms and conditions of the Institute Policy and Intellectual Property (hereinafter Policy) currently in effect, as approved by the competent authority of the Institute.

5. I agree to allow the Institute to make available the abstract of my thesis in both hard copies (printed) and electronic form.

6. For the Institute's own, non commercial, academic use I grant to the Institute the non-exclusive license to make limited copies of my thesis in whole or in part and to loan such copies at the Institute's discretion to academic persons and bodies approved of from time to time by the Institute for non-commercial academic use. All usage under this clause will be governed by the relevant fair use provisions in the Policy and by the Indian Copyright Act in force at the time of submission of the thesis.

7. Furthermore,
    a. I agree to allow the Institute to publish such copies of the electronic version of my thesis on the private Intranet maintained by the Institute for its own academic community.
    b. I agree to allow the Institute to publish such copies of the electronic version of my thesis on a public access website on the Internet should it so desire.

8. That in keeping with the said Policy of the Institute I agree to assign to the Institute (or its Designee/s) according to the following categories all rights in inventions, discoveries or rights of patent and/or similar property rights derived from my thesis where my thesis has been completed.

a. With use of Institute-supported resources as defined by the Policy and revisions thereof,
b. With support, in part or whole, from a sponsored project or program, vide clause 6(m) of the Policy.

I further recognize that:

c. All rights in intellectual property described in my thesis where my work does not qualify under sub-clauses 8(a) and/or 8(b) remain with me.

9. The Institute will evaluate my thesis under clause 6(b1) of the Policy. If intellectual property described in my thesis qualifies under clause 6(b1) (ii) as Institute-owned intellectual property, the Institute will proceed for commercialization of the property under clause 6(b4) of the policy. I agree to maintain confidentiality as per clause 6(b4) of the Policy.

10. If the Institute does not wish to file a patent based on my thesis, and it is my opinion that my thesis describes patentable intellectual property to which I wish to restrict access, I agree to notify the Institute to that effect. In such a case no part of my thesis may be disclosed by the Institute to any person(s) without my written authorization for one year after the date of submission of the thesis or the period necessary for sealing the patent, whichever is earlier.

Somesh Jain
(Name of the student)

Prof. Niloy Ganguly
(Name of Supervisor)

(Signature of the Student)

Department: Computer Science And Engineering

Signature of the Head of the Department

Certificate

This is to certify that the project thesis entitled "**Short Text Classification in Twitter through a Semantic Transform based on Wikipedia**", submitted by **Mr. Somesh Jain**, Undergraduate Student, in the *Department of Computer Science and Engineering*, *Indian Institute of Technology Kharagpur, India*, in fulfilment of the requirements for the degree of **Master of Technology,** is a bona fide record of an original research work carried out by him under the supervision of Prof. Niloy Ganguly (Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur).

Place:
Date:

Prof. Niloy Ganguly
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur

Acknowledgement

I take this opportunity to express my sincerest gratitude to Prof. Niloy Ganguly, Department of Computer Science and Engineering, IIT Kharagpur for being my mentor throughout the course of my thesis work. Without his esteemed guidance, supervision, and thoughtful criticism, construction of this work would not have been remotely achievable. Not only did he provide an insight into the domains of Online Social Networking Analysis and Complex Networks, but also provided enough opportunity and motivation to appreciate the importance of this topic and carry out work on this project with freedom.

I also wish to thank Mr. Parantapa Bhattacharya, PhD scholar, Department of Computer Science and Engineering, IIT Kharagpur and Mr. Saptarshi Ghosh, PhD scholar, Department of Computer Science and Engineering, IIT Kharagpur for providing guidance and helpful advices whenever these were needed. Also, I thank Yahoo India for the summer internship where I worked with problems which had a huge impact on the project.

Somesh Jain
08CS3013
5th year Dual Degree Student
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur

Abstract

Twitter, a micro-blogging service, provides users with a frame-work for writing brief, unconventional, often-noisy postings about their lives. These posts are called Tweets. Currently, millions of Twitter users post millions of 140-character tweets, about topics ranging from daily activities, to opinions, to links, to pictures, etc. The users may become overwhelmed by such huge collection of data covering a broad range of topics. One solution is the classification of short text messages i.e. associating some topics to a tweet. As these short texts are unconventional and do not provide sufficient word occurrences, so traditional classification methods such as 'Bag of Words', Latent Semantic Analysis, Latent Dirichlet Allocation etc. do not fare well. To address this problem, a solution is proposed which uses available knowledge corpus such as Wikipedia, Wordnet etc. for creating a one-to-many mapping between a tweet and Yahoo Content Taxonomy which is a subject ontology. This approach leverages Wikipedia as a knowledge base to disambiguate and categorize the Tweets.

Twitter users are then classified into certain subject categories based on the categories obtained through classification of the user's tweets. Various direct applications of the research are discussed at length to emphasize on the motivation and importance.

This thesis provides insight into all the work done till now, algorithms implemented and the literature read that has been used in conducting out the research. Various tables and figures have also been included with the supporting statistical data in order to provide a clear picture.

CONTENTS

Chapter 1

Introduction

Twitter is an online social networking and microblogging service that enables its users to send and read text-based posts of up to 140 characters, informally known as "tweets". A generic profile on Twitter consists of three major components: the account's tweets, followers, and friends.

Tweets: A tweet is a colloquialism used by Twitter to describe a status update, analogous to an email's body.

Followers: An account's followers are the set of users that will receive a tweet once it is posted, akin to the To field of an email.

Friends: Relationships in Twitter are not bidirectional, meaning a user can receive tweets from a friend without revealing their own tweets. Friends are the set of users an account subscribes to in order to obtain access to status updates.

Hashtags: A hashtag is a word or a phrase prefixed with the symbol #, a form of metadata tag. Short messages may be tagged by including one or more with multiple words concatenated. Hashtags provide a means of grouping such messages, since one can search for the hashtag and get the set of messages that contain it.

Retweets: To repost another user's message on the social networking website Twitter.

Figure 1.1 shows the twitter profile of Barack Obama and all the components present in a profile. Figure 1.2 shows structure of a tweet and features associated with a tweet.

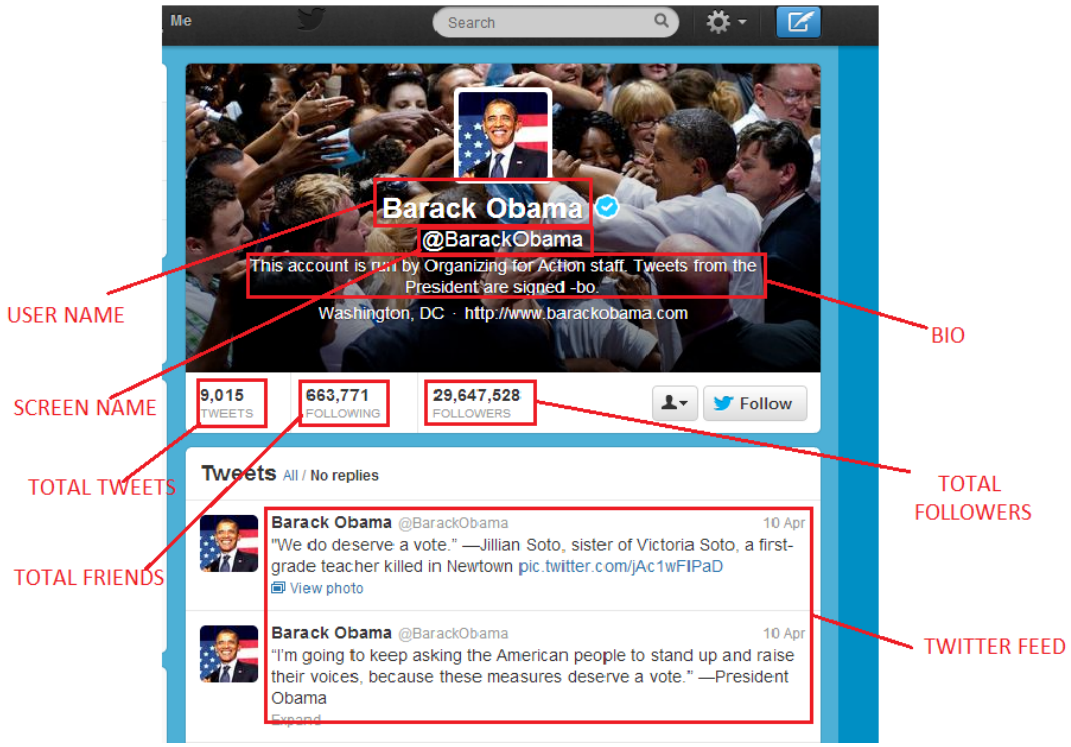

Figure1.1: Twitter Profile of United States President. Barack Obama

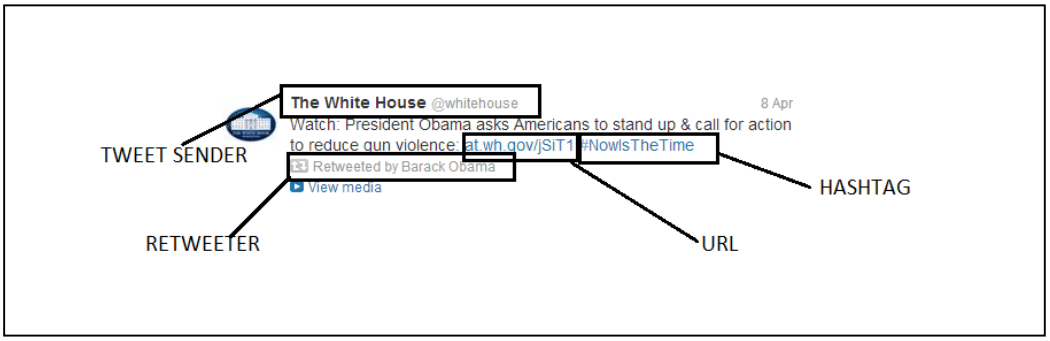

Figure1.2: A tweet and its details

Twitter has grown phenomenally in the last few years with huge amounts of information being generated. Users may become overwhelmed by such huge collection of data. The *tweets* from users are referred to as microblogs because there is a 140 character limit imposed by Twitter for every tweet. This lets the users present any information innovatively and creatively with only a few words, using acronyms and short text optionally followed with a link to a more detailed source of information. Users rely on common acronyms (e.g. “d/r” means “dressing room" in sports), disambiguation via context (“Obama” in a Tweet refers to Michelle Obama instead of Barack Obama), and other constraining mechanisms. However, Tweets can also be information rich, because users tend to pack substantial meaning into the short space.

The first goal of our work is to automatically classify incoming tweets into different topic categories so that users are not overwhelmed by the raw data. . As these short texts are unconventional and do not provide sufficient word occurrences, so traditional classification methods such as 'Bag of Words', Latent Semantic Analysis, Latent Dirichlet Allocation etc. do not fare well.

Our approach to associating topics with the tweet hinges upon finding the entities in the tweet, and then determining a common set of high-level subject categories that covers these entities. Subject categories here refer to the categories which are related to human interests. Eg: Politics, Sports, Entertainment, News etc. Finding named entities and the subject categories is the part where a knowledge base is required because 'Bag of Words' method does not work here as the tweets are short and not in ideal English. This is the part where Wikipedia and its ontology are used.

We develop a technique for mapping a tweet to Yahoo Content Taxonomy using Wikipedia structure and information. The next subsections discuss the structures of Wikipedia and Yahoo Content Taxonomy. These two are the integral parts of the research. Figure 1.3 shows the framework of the model.

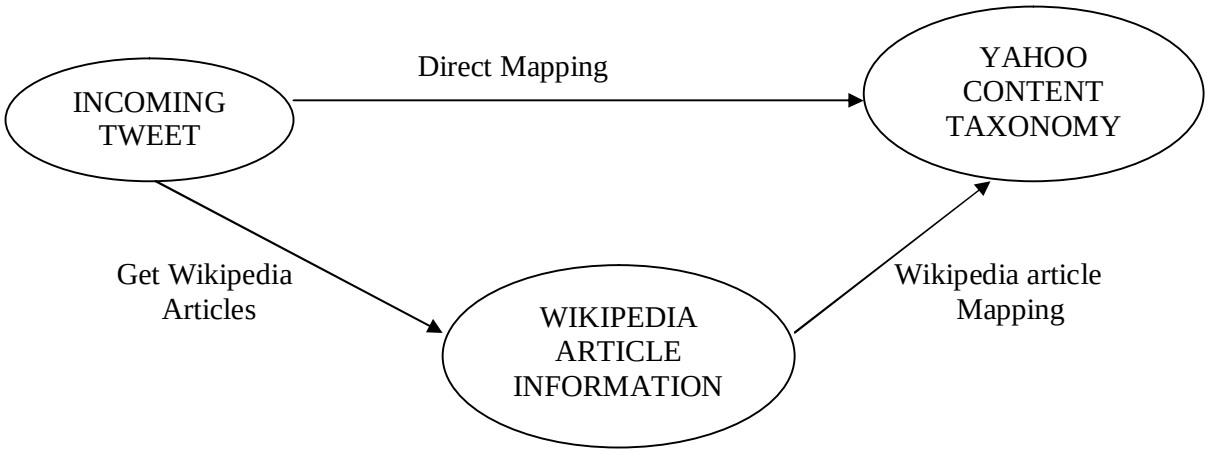

Figure 1.3 : Framework of the research

1.1 Wikipedia and its Structure

Wikipedia is a free, collaboratively edited, and multilingual Internet encyclopaedia supported by the non-profit Wikimedia Foundation. Its 23 million articles, over 4 million in the English Wikipedia alone, have been written collaboratively by volunteers around the world. In this research, *only the articles written in English language are considered because we take only those tweets into account which have been posted in English.*

Wikipedia is an encyclopaedia where articles are not independent but are related to each other through a taxonomy. Every Wikipedia article is the child of various parent categories. Wikipedia categories are the inner nodes of the Wikipedia taxonomy where every Wikipedia article is connected to at least one category which is considered as the article's parent category. These categories further have different categories as their parent categories and this is the way the ontology is constructed. *So, every category could have more than one subcategories, many parent categories and many articles in the category.* Figure 1.4 and 1.5 shows some of the parent categories of Barack Obama and pages in a category.

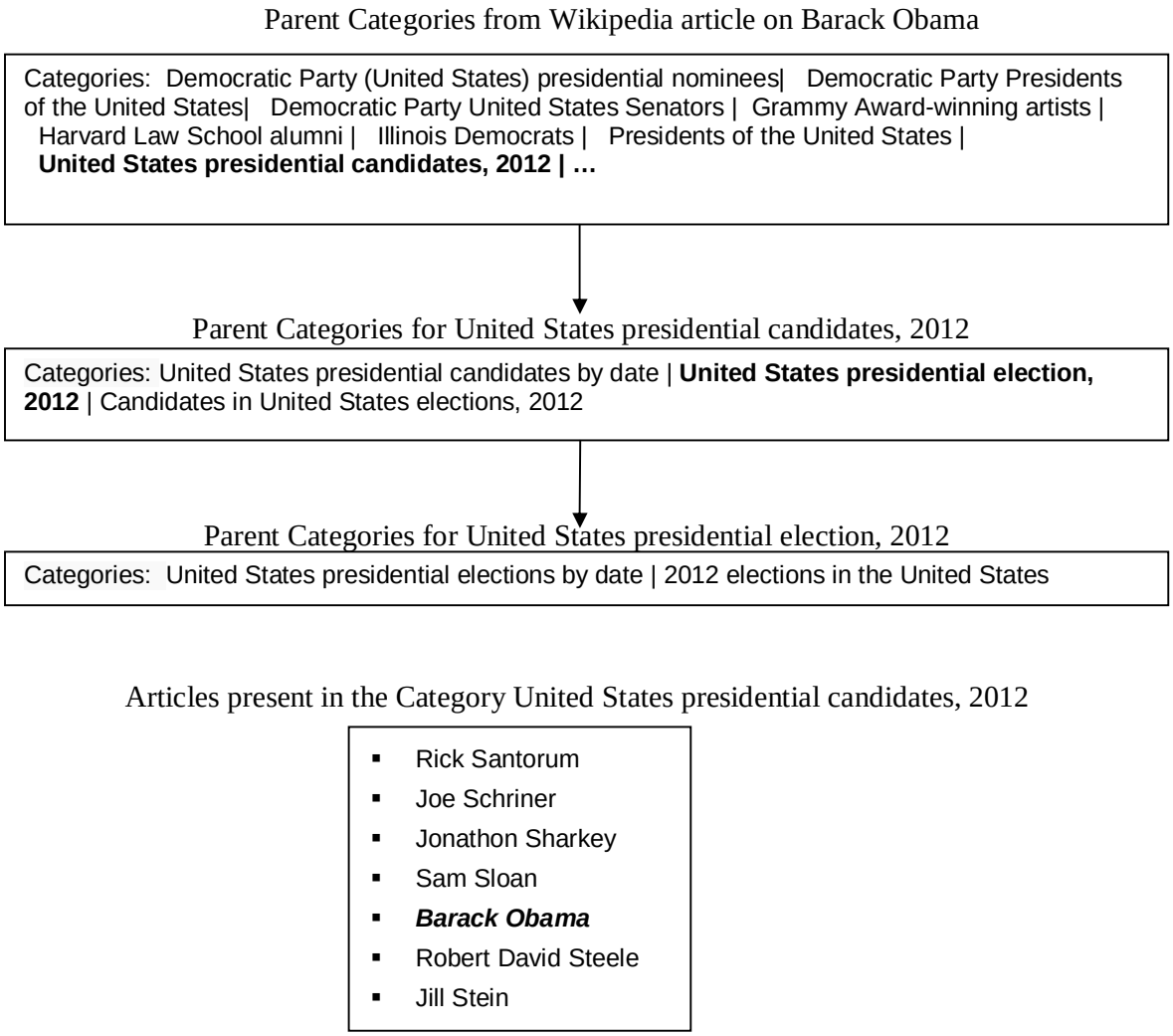

Figure 1.4 : Parent categories of Wikipedia article on American incumbent President Barack Obama

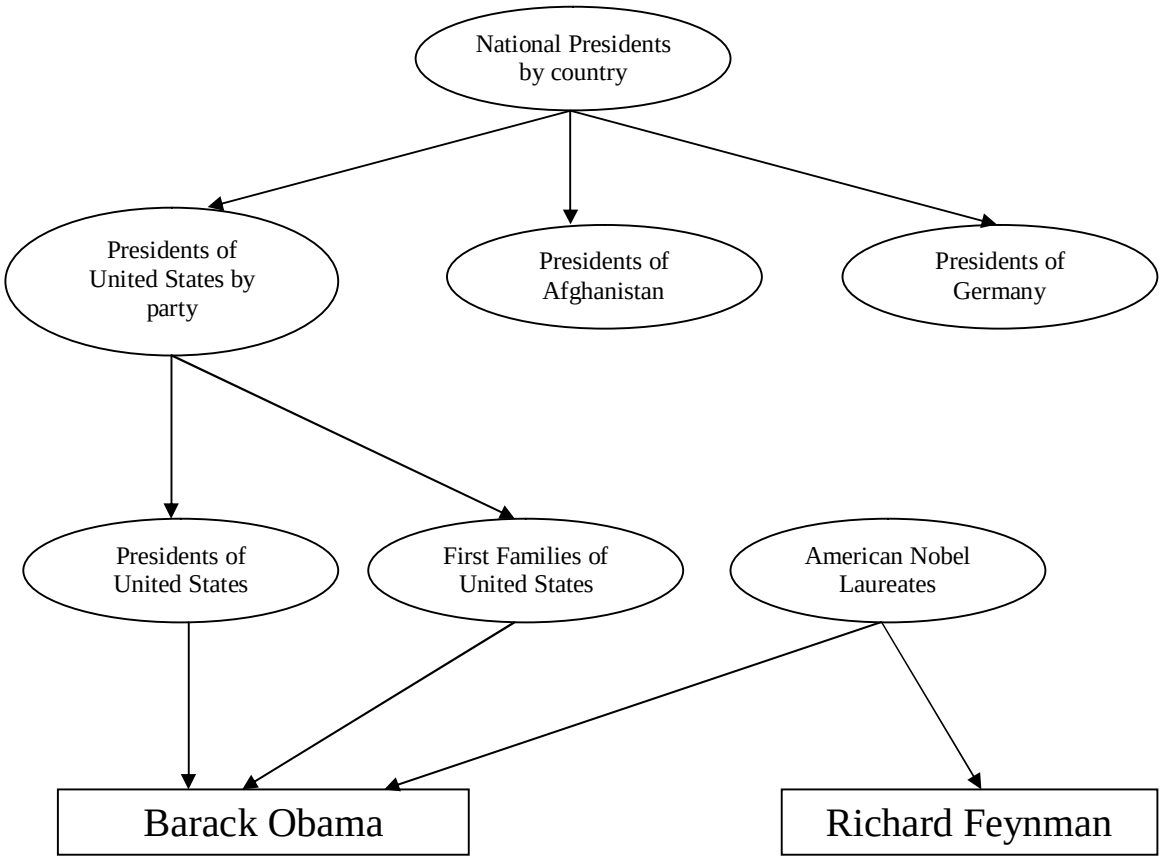

Figure 1.5 Part of folksonomy tree of Barack Obama

Some Wikipedia facts:

1. Number of English language articles --- 4 million
2. Number of English language categories --- 0.7 million

The above specified structure of Wikipedia articles and their parent categories is heavily used in the proposed approach which is discussed later.

1.2 Yahoo Content Taxonomy (YCT)

Yahoo Content Taxonomy is an ontology which is divided into 7 dimensions. In this project, we are interested only in the subject dimension because it covers interests and hobbies. Around 1000 categories are present in 4 levels in the subject dimension.

Wikipedia taxonomy is an acyclic graph where every category could have more than 1 parent categories and can have more than 1 pages present in the category. Unlike Wikipedia taxonomy, Yahoo Content Taxonomy is a hierarchical tree structure with depth of 4 levels.
Eg: : /Sports & Recreation/American Football/

Figure 1.6 show dimensions *people* and *regions* in yct. Figure 1.7 shows the top level category names in yct - *subject* dimension. Figure 1.8 shows a subtree of a top level *subject* dimension category – *Sports & Recreation, Politics & Government*.

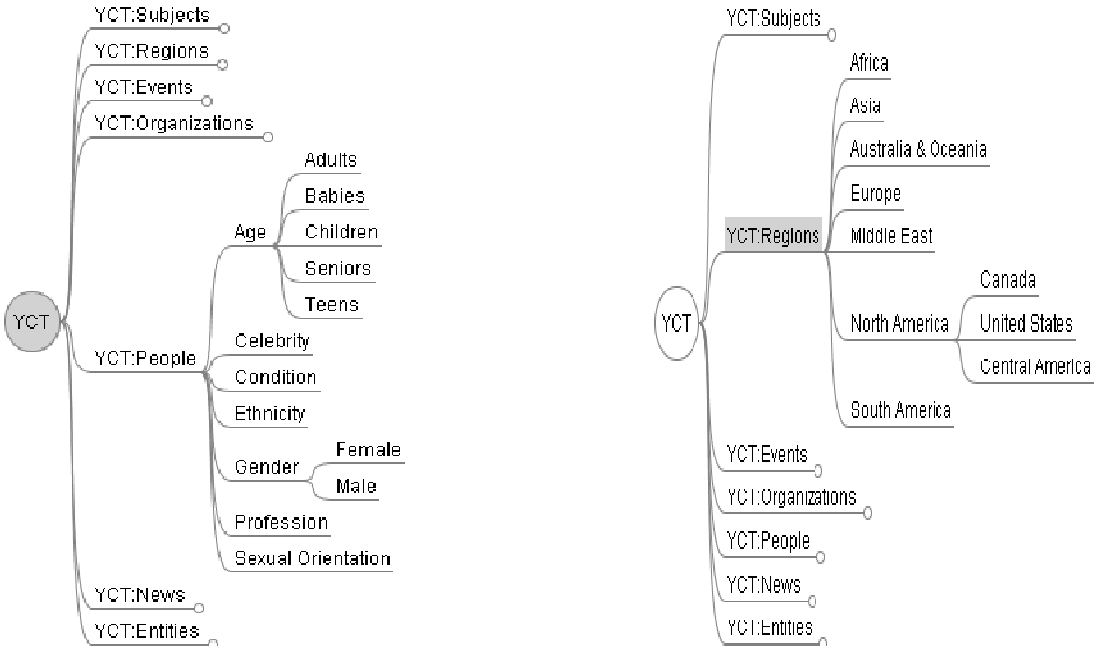

Figure 1.6 Yahoo Content Taxonomy dimensions : People, Regions

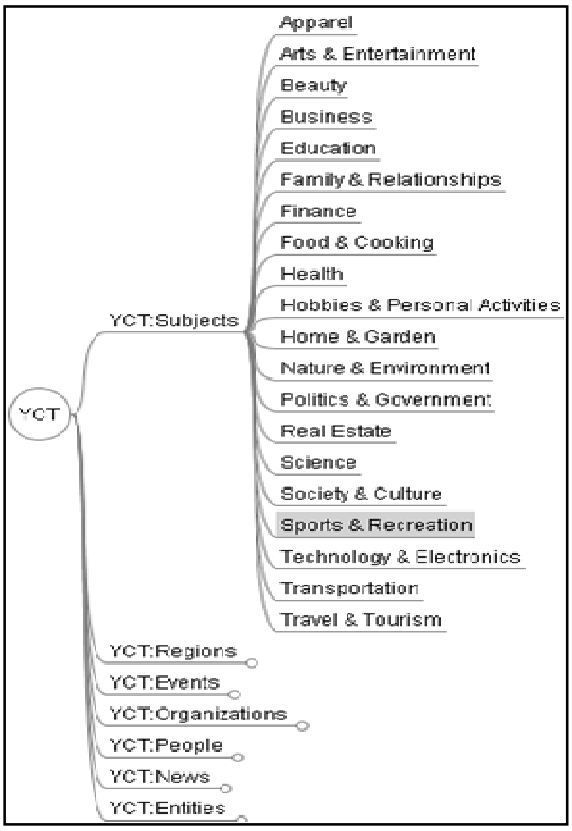

Figure 1.7 Top level categories of YCT Subject dimension

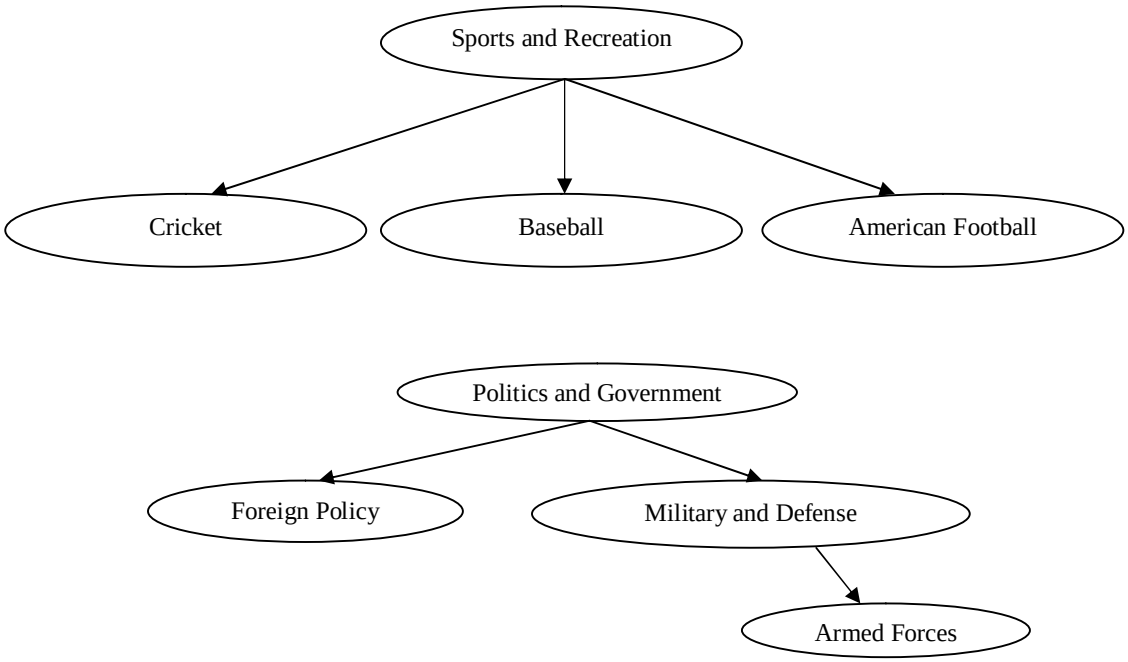

*Figure 1.8 Part of subtrees of **Sports and Recreation**, **Politics and Government***

Chapter 2

Literature Review

As stated in the proposal, the primary goal here is to automatically classify incoming tweets into different categories. The tweets (micro-blogs) are considered as short text messages and following classification schemes have been tested.

Classifying tweets is not an easy task because statistical methods of text classification have difficulty on short texts [1]. Moreover, if emerging topics of conversation are regarded as signal, the vast majority of tweets would be characterized as noise. Some past studies of tweet classification have examined the use of specific features, such as emoticons [2] and author profiles [3], in improving the classification performance. Other studies have regarded tweets as a window into customer perception [4]; then, the challenge becomes recognizing sentiment. Various authors have attempted to "give meaning" to text in general[13] or to text contained in tweets. Liu et al. [10] focus on NER on tweets and use a semi-supervised learning framework to identify four types of entities. Benson et al. [11] try to match tweets to "records." These records are, for example, artist-venue pairs and can be obtained from sources like music guides. They train a model that extracts artists and venues from tweets and automatically match these to the extracted records.

In contrast to these past works, we are interested in categorizing tweets in order to detect topics, which requires the ability to cluster tweets without a priori knowing which features will be important. Recent work on extracting topics from short texts relies on knowledge bases to find context that is not in the texts. For example, Stone et al. used Wikipedia as training corpus to improve the ability of statistical methods to discover meanings of short texts[5]. Existing works on classification of short text messages integrate messages with meta-information from other information sources such as Wikipedia and WordNet [6,7]. Sankaranarayanan et al [8] introduce TweetStand to classify tweets as news and non-news. Automatic text classification and hidden topic extraction [1] approaches perform well when there is meta-information or the context of the short text is extended with knowledge extracted using large collections. Our approach is more general when compared with the TweetStand. Sriram et al classifies incoming tweets into categories such as News (N), Events (E), Opinions (O), Deals (D), and Private Messages (PM) based on the author information and features within the tweets.

Similarly, Gabrilovich and Markovitch used concepts derived from Wikipedia to identify the semantic relatedness of texts [9]. Wikipedia was also used by Michelson and Macskassy [9]. Meij et al. propose a solution to the problem of determining what a microblog post is about through semantic linking. They add semantics to posts by automatically identifying concepts that are semantically related to it and generating links to the corresponding Wikipedia articles.

The solution proposed here is semi-supervised. More information is associated with the post using the information already present in the tweet and information extracted from Wikipedia regarding the entities in the tweet. The solution uses an already existing named entity recognition technique to get the named entities present in Wikipedia related to the tweet and use the information provided by Wikipedia to classify the tweet into an already existing set of categories present in Yahoo Content Taxonomy. The solution is discussed in the next section.

Genc et.al[12] classify tweets using Wikipedia semantic knowledge and compare the results with results of Latent Semantic Analysis (LSA) and String Edit Distance (SED) which support that Wikipedia approach provides better results in case of tweet classification. There has been almost no technique to my knowledge where a predefined elaborate set of user interests is taken as classification categories.

So, most of the approaches that have been used for tweet classification do classification without incorporating knowledge from any knowledge base. We use Wikipedia as external knowledge base for classification. Also, models previously used rely heavily on training the model through a set of tweets but that makes the classification scheme dependent on a specific set of words, acronyms etc. Here, we are taking any training data for our classification. Previous models have taken very few and most broad categories for classification such as news, non-news, private messages etc. Our proposed model takes tweets and classifies them into specific user interests which increase the applications of tweet classification.

Chapter 3

Tweet Classification Technique

Every incoming tweet is classified into subject categories present in the Yahoo Content Taxonomy using Wikipedia taxonomy information and Yahoo Content Analysis API. Yahoo Content Analysis is the named entity recognition technique that has been used in the proposed solution. It is discussed in the next subsection. Figure 3.1 shows the approach of the tweet classification model to be discussed in this chapter.

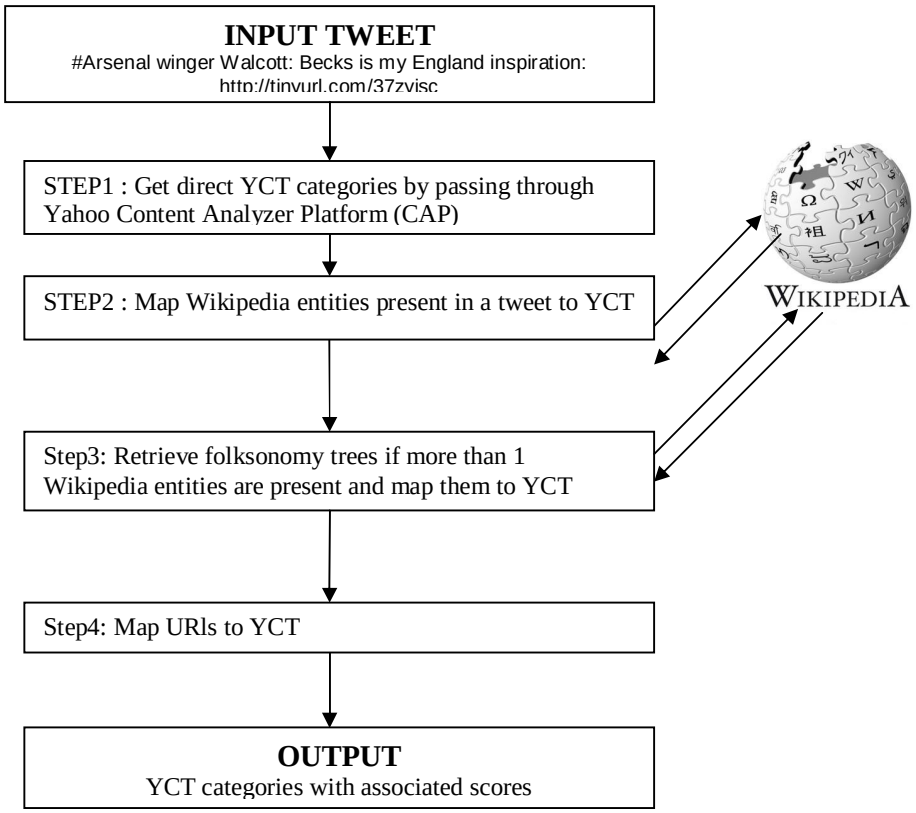

Figure3.1 : Approach of the model proposed for a single tweet classification

3.1 Yahoo Content Analysis Platform

This is the first step in the solution discussed. The tweet is passed through Yahoo Content Analysis Platform (CAP) which provided named entities present in the tweet.

CAP web service detects entities/ concepts, categories, and relationships within unstructured text. It ranks those detected entities/concepts by their overall relevance, resolves those if possible into Wikipedia pages, and annotates tags with relevant meta-data. The

Content Analysis service is available as a Yahoo Query Language (YQL) table. CAP search works on the Bag of Words model. This makes our classification technique semi-supervised.

Parameters required : Text present in the tweet (Text includes hashtags, plain text and URL links).

Response fields: Response fields refer to the structured output obtained when tweet is passed through CAP .

There are various response fields such as *categories/yct_categories/yct_category , entities/entity/text, entities/entity/wiki_url, entities/entity/types, entities/entity/types/type, entities/entity/metadata_list,entities/entity/metadata_list/metadata, entities/entity/metadata_list/metadata/geo_area,entities/entity/related_entities/Wikipedia, entities/entity/related_entities/wikipedia/wiki_url* etc. For the project, we have used only a few response fields such as *categories/yct_categories/yct_category, entities/entity/wiki_url*

categories/yct_categories/yct_category	YCT category. This element has a numeric score attribute. Categories are listed in descending order of scores. Range of scores is [0,1] but are not normalized.
entities/entity/wiki_url	The Wikipedia URL of the entity/concept. Not present when the entity/concept doesn't have a Wikipedia page.

Response format is shown in Figure3.2 for a particular tweet :

Manmohan Singh is the prime minister of india.

Figure3.2 Response fields of CAP results for the given tweet

```
<?xml version="1.0" encoding="UTF-8"?>
<query xmlns:yahoo="http://www.yahooapis.com/v1/base.rng"
   yahoo:count="2" yahoo:created="2013-04-07T15:46:39Z" yahoo:lang="en-US">
   <diagnostics>
      <publiclyCallable>true</publiclyCallable>
      <user-time>18</user-time>
      <service-time>15</service-time>
      <build-version>35405</build-version>
   </diagnostics>
   <results>
      <yctCategories xmlns="urn:yahoo:cap">
         <yctCategory score="0.694444">Executive Branch</yctCategory>
         <yctCategory score="0.590597">Politics &amp;amp; Government</yctCategory>
      </yctCategories>
      <entities xmlns="urn:yahoo:cap">
         <entity score="0.963">
            <text end="13" endchar="13" start="0" startchar="0">Manmohan Singh</text>
            <wiki_url>http://en.wikipedia.com/wiki/Manmohan_Singh</wiki_url>
            <types>
               <type region="us">/person</type>
               <type region="us">/person/director</type>
               <type region="us">/person/government/world_leader</type>
            </types>
         </entity>
      </entities>
   </results>
</query>
```

So, every incoming tweet is passed through CAP and yctCategories list and wiki url lists are obtained if present. As specified earlier, the first task is to map every tweet to yct categories if possible. There are 2 possible cases :

Case1 : CAP itself maps the tweet to some yct categories with their respective scores as is done for the tweet shown in the above figure . For a tweet to have some yct mapping available directly from Yahoo Content Analysis search, the tweet has to be less innovative. In this context, less innovativeness means that the tweet is written in plain English with least use of acronyms, abbreviations etc. It can be inferred that such tweets are very less because people try to post a tweet creatively using acronyms etc. so that they can post more and more information in the 140 character limit.

Since most of the tweets do not have a direct yct classification as specified in case 1 discussed, case2 is used for tweets which have some associated wiki entity present in the CAP result. We'll see later in the results and analysis section that many tweets have wiki entities present in the response field of CAP search.

Case2 : Response fields in CAP search give Wikipedia entities present in the tweet. More than one Wikipedia entity could be present in the CAP result. Most of the sensible tweets i.e. the tweets which could be classified into certain topics have some Wikipedia entities present in them. So, if CAP provides us with these wiki entities (names of Wikipedia articles), then a mapping could be generated between Wikipedia article and Yahoo Content Taxonomy and so, *a tweet can be classified using the classification of all the Wikipedia entities present in the tweet* .

This mapping of a Wikipedia article to YCT categories is non – trivial and we propose an algorithm which maps every Wikipedia article to yahoo content taxonomy. This separate model is then used in the project for tweets where Wikipedia entities are present in the CAP results of the tweet.

We tackle the cases of exactly one Wikipedia article and more than one Wikipedia articles differently. Firstly, the former case is discussed and then the latter one.

3.2 Map Wikipedia article to Yahoo Content Taxonomy

This problem is discussed here separately. Solution is proposed and then this model is merged with our existing model to classify incoming tweets using CAP search . We have different solutions for the cases when there is exactly one Wikipedia article and for the cases when there are more than one Wikipedia articles.

This differentiation is trivial to understand. If there is only one Wikipedia article, then the best mapping would be to map the article according to the most important things pertaining to that article. For eg: Wikipedia article on *Barack Obama* should mapped to *American Presidents, Democratic Party Presidents* rather than to *Nobel Laureates, Illinois Democrats* as the former categories are better than the latter to categorize him. When there are more than one Wikipedia entity, then our aim is not to categorize each of them as separate entities but to categorize the whole as a single entity. So, what it means is that we are more

interested in getting the relation which is common between the articles. For instance, if Barack Obama and Mitt Romney are the articles in question, then we would like the best subject categories as *United States Presidential Candidates, 2012* etc. This is the reason why both the cases have been handled separately.

In the next subsection 3.2.1, mapping technique for a single Wikipedia article is discussed and in subsection 3.2.2, mapping technique for more than one Wikipedia articles is discussed.

3.2.1 Mapping a Wikipedia article to Yahoo Content Taxonomy

In section 3.1, we discussed that the mapping of a Wikipedia article to YCT categories will be used in our main model of mapping tweets to YCT categories. Here in this section, mapping of a single Wikipedia article to YCT is discussed.

Framework of the approach is described in Figure3.3

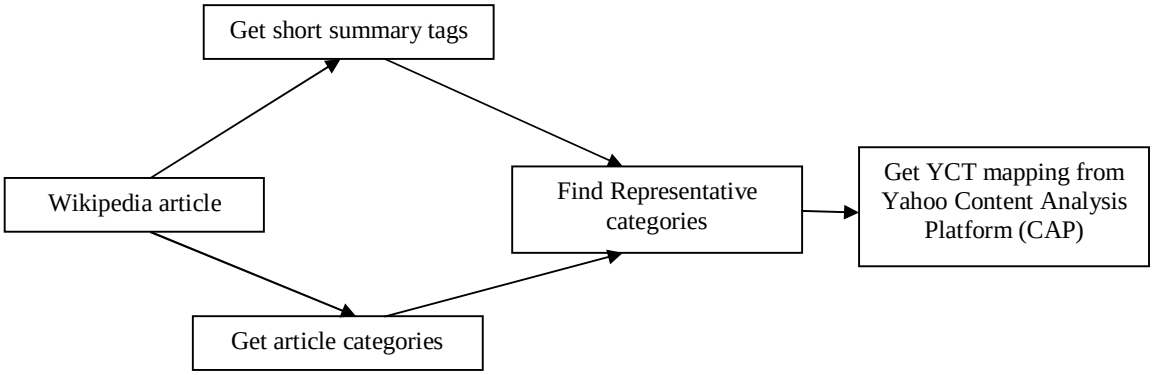

Figure3.3 : Framework for mapping Wikipedia article to YCT

All the information that we extract for a Wikipedia article is done using DBpedia which provides structured information. We query DBpedia using SPARQL for short abstract summary of the article and for parent categories of the article.

Steps involved in the process are described here :

Step1 : Get short summary tags

Short summary refers to the text which is present on the top of any Wikipedia article which can be considered as the most important facts regarding that article. Short summary of Wikipedia article on Barack Obama is given in figure. Short summary is extracted using SPARQL query on DBpedia. Figure 3.4 shows abstract summary of Wiki article on Barack Obama

> Barack Hussein Obama II is the 44th and current President of the United States. He is the first African American to hold the office. In January 2005, Obama was sworn in as a U.S. Senator in the state of Illinois. He would hold this office until November 2008, when he resigned following his victory in the 2008 presidential election. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was the president of the Harvard Law Review.

Figure3.4 : Short Abstract summary of Wikipedia article on Barack Obama

Step 2 : Get Article Parent Categories

Section 1.1 discusses the Wikipedia article and its parent categories. Here, we extract only the names of level1 parents i.e. direct parent categories of the article. Again, SPARQL query on DBpedia is used for obtaining parent categories. Figure 3.5 shows first level parents of Wikipedia article on Barack Obama

Democratic Party (United States) presidential nominees | Democratic Party Presidents of the United States | Democratic Party United States Senators | Grammy Award-winning artists | Harvard Law School alumni | Illinois Democrats | Illinois lawyers Illinois State Senators | Nobel Peace Prize laureates | Obama family | Occidental College alumni | People from Honolulu, Hawaii | Politicians from Chicago, Illinois | Presidents of the United States | Punahou School alumni | United Church of Christ members | United States presidential candidates, 2008 | United States Presidential Candidates, 2012| United States Senators from Illinois | University of Chicago Law School faculty | Writers from Chicago, Illinois |

Figure3.5: Parent Categories in Wikipedia taxonomy for Barack Obama

Step 3: Representative Categories

In the above figure, it can be observed that there can be many categories which are correct for the article but are not the most important categories to represent that article. For eg: In case of *Barack Obama, Grammy Award-winning artists, United Church of Christ members, Punahou School alumni, University of Chicago Law School faculty* etc. are not important. The most important ones are *Presidents of the United States, Democratic Party Presidents of the United States, United States Presidential Candidates, 2012 etc.* We call these categories as the representative categories.

The aim of this step is to select the representative categories from the set of article categories obtained in step 2. This is done using the short summary obtained in step1.

The short summary is speech tagged and we obtain the set of nouns, groups of proper nouns. This group of nouns, proper nouns is called as short summary tags. After this, wordnet ontology which is an ontology of concepts is used to associate synonyms of the already obtained short summary tags. These synonyms are also added to the list of summary summary tags. We do not take the complete short summary but just the first line of the short summary.

In case of *Barack Obama* , tags are – *President, United States* .

From the list of article categories obtained in step 2, only those categories are selected which have maximum match with the list of short summary tags. Here, match means that the article name contains a word which is present in the list of tags. Figure 3.6 shows representative categories obtained after this step

Democratic Party (United States) presidential nominees | Presidents of the United States | Democratic Party Presidents of the United States | Democratic Party United States Senators | United States presidential candidates, 2008 | United States Presidential Candidates, 2012|

Figure3.6 : Representative categories for Barack Obama

While comparing the tags, we filter out English stopwords and stem the words using porter's stemming algorithm.

Step 4: Feed Representative categories to Contextual Analysis Platform (CAP)

Contextual Analysis Platform has been discussed thoroughly in Section 3.1. Here, we form a string of the representative categories obtained in the last step. We take a maximum of 5 category names sorted by their matching scores with the summary tags.
For Barack Obama, the string input given to CAP is : *Presidents of the United States | United States presidential candidates, 2008 |Democratic Party (United States) presidential nominees | United States Presidential Candidates, 2012| Democratic Party Presidents of the United States | Democratic Party United States Senators*

CAP returns the output in similar fashion as it return for the case of a tweet. So, we just consider this formed string as a tweet derived from the original tweet and feed it to CAP. Unlike our main method, where we extract yctcategories and wiki entities from the response fields of CAP, here we consider only yctcategories results.

Some examples of mapping are in Figure3.7

Barack_Obama
['Executive Branch', '0.955556']
['Politics & Government', '0.941406']
['Elections', '0.875']

Winston_Churchill
['Politics & Government', '/Politics & Government/', '0.98']
['Science', '/Science/', '0.848837']
['History', '/Science/History/', '0.64']
['Armed Forces', '/Politics & Government/Military & Defense/Armed Forces/', '0.5']

2010_Haiti_earthquake
['Society & Culture', '/Society & Culture/', '0.961165']
['Disasters & Accidents', '/Society & Culture/Disasters & Accidents/', '0.873684']
['Natural Phenomena', '/Nature & Environment/Natural Phenomena/', '0.529412']

Country_music
['Arts & Entertainment', '/Arts & Entertainment/', '0.952135']
['Music', '/Arts & Entertainment/Media/Music/', '0.929577']
['Media', '/Arts & Entertainment/Media/', '0.885714']

Figure3.7 : YCT mappings with associated scores of certain Wikipedia articles

Here, several other approaches were tried out. One approach was to feed the entire category list to CAP. Other approach was to use the complete summary and not just the short summary. Third approach was without the use of wordnet. In the results section, we discuss the results of these approaches and their limitations. For our project of mapping a tweet to yct

categories, we use the above described approach as it performs better than the other mentioned techniques.

3.2.2 Mapping more than one Wikipedia article to Yahoo Content Taxonomy

CAP might give more than one Wikipedia entity i.e. more than one Wikipedia article when a tweet is fed to it. In such a case, what is needed is not the best categories of all the articles but the best common categories between these articles.

For example: Consider a tweet – *Both Obama and Romney are good choices.*

When this tweet is fed, we don't need *American Presidents, Businessman, Illinois Senator*s etc. as the topmost categories because these categories are reflective of either Obama or Romney but not of both What is needed is categories like *United States Presidential Candidates, 2012 , American Politicians etc.* because these categories represent both the articles.

It is clear that approach discussed in section 3.2.1 cannot be used here. The algorithm used here takes use of Wikipedia folksonomy tree of the articles. In the discussion and example described here, we are considering only 2 articles but the approach can be generalized for more than 2 articles also. Here, we extract parent categories of a Wikipedia page up to 3 levels. This means that we consider direct parent categories of the article, parents of these direct categories and their parents. Extraction is done using SPARQ query on DBpedia. Query extracts names of categories upto 3 levels with the connections between these categories.

Part of Wikipedia category trees upto 3 levels are shown for Barack Obama and Mitt Romney in Figure 3.8. Only 3 levels of Wikipedia ontology is considered in our case. This is because after 3 levels, categories become too generic and confusing. In the following figures, the common categories between the 2 articles are shown in bold. These categories shown in bold are the representative categories for this case of mapping more than one Wikipedia entities. Note that these representative categories are spread over 3 levels as compared to the representative categories defined in section 3.2.1 which belonged only to level1.

Only the top 5 common categories are selected to be fed to CAP. We associate scores with the categories using the following mathematical formula.

To rank the categories, scoring is performed in the following manner:
Let a category is represented by 'c'. Score is associated in this way:

$$Score(c) = Freq(c) \times w_c$$

Freq(c) refers to the frequency of category 'c' occurrence.
w_c refers to the weight associated with category 'c'. In the weight equation, the level of the category is taken into account. The category which is closer to the node(entity) has to be given more score. So, weight wc is inversely proportional to the category's level.

$$w_c = (1/2)^d$$

Here, 2 is taken as the factor. Instead of 2, branching factor in Wikipedia ontology could also be taken.

Inconsistency might arrive when same category occur in different levels in different entity trees. To avoid this, we take the ranking score of a category as the sum of its ranking scores for each depth in each tree where it occurs.

For instance, if a category occurs 3 times in 2^{nd} level , 2 times in 3^{rd} level, then:

Score on 2^{nd} level = $3x(1/2)^2 = ¾$

Score on 3^{rd} level = $2x(1/2)^3 = ¼$

Total ranking score = Score on 1^{st} level + Score on 2^{nd} level = ¾ + ¼ = 1

At the end of this step, categories have some associated ranking scores. The category with the highest score means that category might have been common in more than 1 Wikipedia entity folksonomies which augments the ranking scores of the category.

Figure 3.8 show the common categories between the 2 Wikipedia articles in bold.

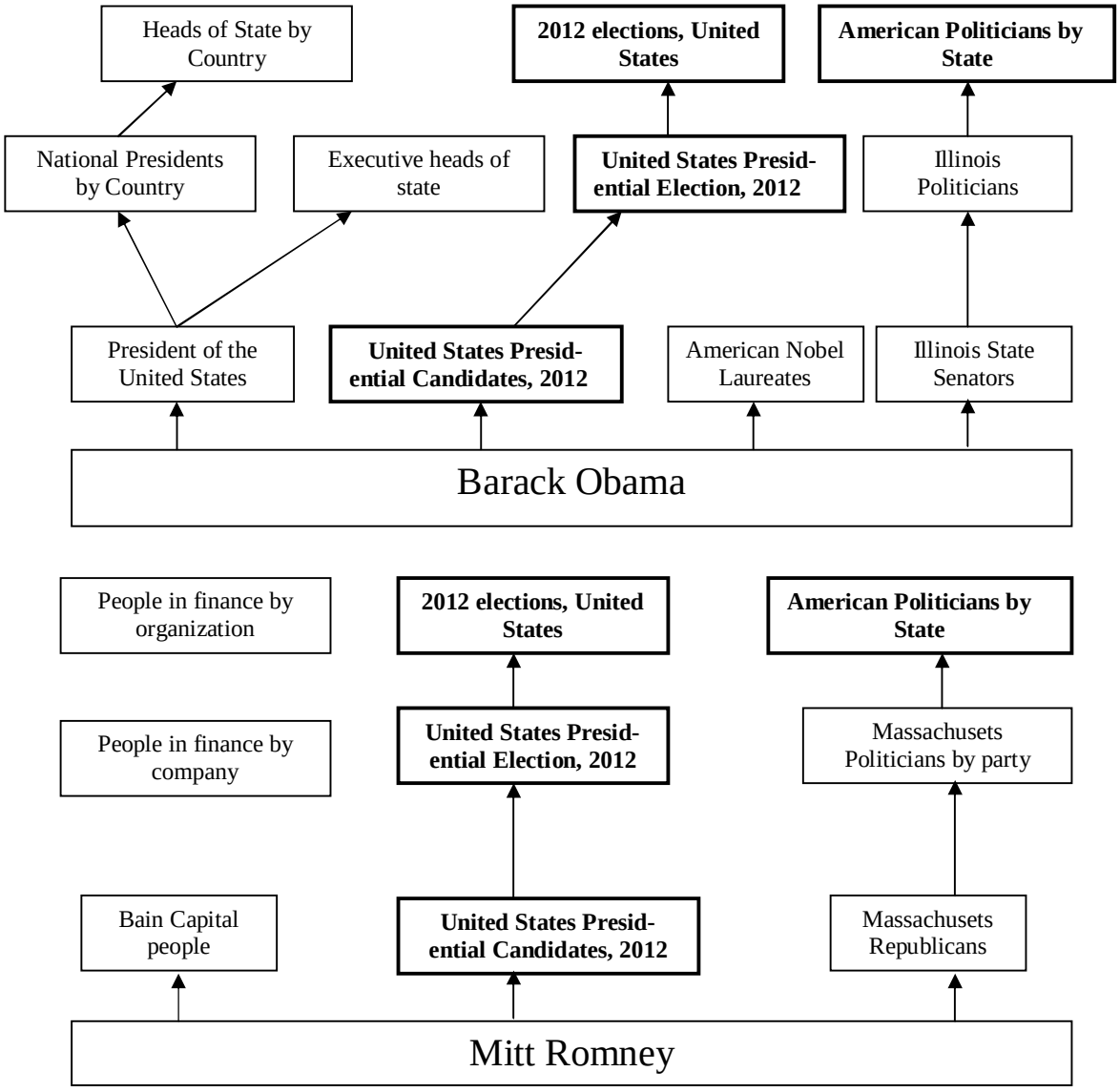

Figure3.8 : Part of folksonomy tree of Wikipedia article on Barack Obama and Mitt Romney

Representative categories are selected according to the scores calculated using the above formula. Top 5 score categories are selected. It is obvious now that these categories would be the best common categories between the set of articles. The set of articles mentioned here can any in number. We have shown it using 2 Wikipedia articles but it can be generalized to any number of articles.

After the representative categories are obtained, a string is formed which is the concatenation of all the obtained categories. This string is fed into CAP search which provides the YCT mapping in the response field *yctmapping*

In case of Barack Obama and Mitt Romney, the representative string is *United States Presidential Candidates 2012 , United States Presidential Election, 2012 elections United States, American politicians by state.* Output obtained by CAP now are yct categories *Politics, Elections* with scores 0.89 and 0.85 respectively.

Here, we have modified the original tweet to a new tweet which is the combination of all the common categories between Wikipedia articles to get yct mapping. The mapping obtained for this modified string is treated as the yct mapping for the original tweet.

The above mentioned techniques discussed in section 3.2.1 and section 3.2.2 are applied after the original tweet is fed to CAP search. Analysis of our integrated algorithms would show that major improvement has occurred after incorporating Wikipedia mapping in the CAP search results.

This ends the tweet mapping algorithm proposed in the research. The above solution works for mapping a single tweet to Yahoo Content Taxonomy. Section 3.3 is an augmentation of the proposed solution. We propose how a user is classified into certain categories based on the classification of his tweets.

3.3 Classification of Tweet URLs

After the implementation of Section 3.2 i.e. getting categories form the Wikipedia pages, we handle the classification of tweet urls.

Because of twitter character limit of 140 characters, most of the twitter users try to post their news item by providing a *tinyurl (t.co) link* regarding the post.
For eg: *Today in Sports Plus: #RedSox http://t.co/2FwOH7bS*

Many tweets contain just the URLs without any supporting text. In order to classify such tweets, we assume that tweet classification of these tweets would be same as URL classification.

A simple technique is used for this classification. For every URL posted, we get the *landing page* of the URL and extract the title of the URL page. Now, we treat this title as a derived tweet and feed it to CAP search. The response results are considered as the results for the tweet classification.

For the above mentioned tweet, landing page is obtained for URL - *http://t.co/2FwOH7bS* . Landing page - *http://myemail.constantcontact.com/Today-in-Sports-Plus---What-moves-should-Red-Sox-make-.html?soid=1105067466523&aid=eeEWZp0kFdk.* Title extracted from the landing page – *What moves should Red Sox make* is considered as a derived tweet and is fed to CAP search which provides new yct response results. In this case, yct categories obtained are *Sports , Baseball.*

The next section discusses the results and analyses the proposed model on various Twitter users and justifies the improvements of the model over the existing techniques.

Chapter 4

Tweeple Classification

User classification means classifying a twitter user into certain subject categories with their associated scores.

Here, yct categories are taken as subject categories. Since, yct categories are around 1000 in number, it can be stated that a user can be classified into some categories out of these 1000 categories.

Classification of a single tweet does not have that many applications as user classification. A few applications of user classification are discussed later. A twitter user might tweet randomly a few times. Here, random means that the user is tweeting about certain topics which are not among his interests. For example, if an exciting cricket match is going on, some user might tweet about it, even if he isn't interested in cricket in general.

For user classification, we assume that if a considerable amount of user's tweets are taken (let's say 100 or 200), then a majority of those tweets would be pertaining to his expertise or knowledge. If this majority of tweets are classified correctly, then a user can be classified correctly.

The classification of users for the purpose of this project has been done using 100 tweets for every user.

For a particular yct category, threshold is selected as 10 tweets i.e. that yct category occurs in the classification of minimum 10 tweets. For every category, scores are added over all the tweets for that category. If minimum support in every tweet i.e. yct category score in a tweet is 0.5, then total minimum score for a category over a span of 100 tweets should be 0.5x10 = 5

Assumptions : minimum threshold for tweets (t_s) = 10
Minimum support for a category in a tweet (c_s) = 0.5

Minimum score required for a yct category to be included in the list of user's expertise categories $(u_s) = t_s \times c_s = 10 \times 0.5 = 5$

Above threshold helps in filtering out all the yct categories which were included in the list of yct categories obtained for 100 tweets of user because they were classified in a few tweets. We select only the categories which cross the threshold score u_s

Figure 4.1 shows tweets of a Twitter user and the classification obtained through the model.

Would you like to finish this sentence? It`s so cold today ...
Today in Sports Plus: Patience in the market #constantcontact #RedSox http://t.co/AJBVclta
Today in Sports Plus: Ben to go all out - or not? #constantcontact #RedSox http://t.co/0wQgGwAe
Today in Sports Plus: What moves should Red Sox make? #constantcontact #RedSox http://t.co/2FwOH7bS
What you think about the Red Sox signing Mike Napoli?-Jerry
Today in Sports Plus: With Napoli deal done, could Josh Hamilton be next? #constantcontact #RedSox http://t.co/K4g7Al50
Free $50 Gift Card with purchase of Four $25 cards to any Jerry Remy`s location #constantcontact http://t.co/ibyo5o7c
This picture came from the archives of the Remy Report. The caption read Big Papie`s Polar Express! http://t.co/psWjvzNs
Today in Sports Plus: #RedSox have character - but what about pitching? #constantcontact http://t.co/5gYSlkWS
Today in Sports Plus: Ellsbury trade talk, and catching up with Francona #constantcontact #RedSox http://t.co/aTVFvhwx
Just found this interesting site all about Matt Cain`s perfect game this year-you should check it out- http://t.co/2tugtb7p #mlb -Jerry
Today let`s remember the men and women who gave their lives at Pearl Harbor-Jerry
Today in Sports Plus: Pedro to front office a possibility for #RedSox #constantcontact http://t.co/Zyy8Fctw
Today in Sports Plus: 2013 #RedSox edition - love `em or hate `em? #constantcontact http://t.co/FWs62nry
Today in Sports Plus: If Lester is to go, it better be for big #constantcontact #RedSox http://t.co/KkOEuET2
Today in Sports Plus: Lester not enough trade bait for KC #constantcontact #RedSox http://t.co/oNoc0hN0
Today in Sports Plus: Yooooouk in pinstripes - done deal #constantcontact #RedSox http://t.co/j7hwbgsx
Ring in the New Year at Jerry Remy`s! #constantcontact #RedSox http://t.co/5OUpgDV9
Today in Sports Plus: What`s the deal, Napoli? #constantcontact #RedSox http://t.co/q40F6pBy
Today in Sports Plus: Mediocrity? #constantcontact #RedSox http://t.co/czYjZ7Ja
Ok I am down to 11 days before Christmas-no idea what to get my wife-some things never change!-Jerry
Today in Sports Plus: #RedSox fans - and Youk - never imagined pinstripes #constantcontact http://t.co/CVml3wBU
Today in Sports Plus: #RedSox staying the course and keeping it safe #constantcontact http://t.co/9ZLJKcIk
Have any Red Sox player stats questions? These guys at http://t.co/aTDYJQG8 can get you an answer fast! http://t.co/aVP4x9nX #askastat
Today in Sports Plus: Another Drew chapter for Red Sox #constantcontact #RedSox http://t.co/32DLjorL
Today in Sports Plus: #RedSox moves - how do they stack up? #constantcontact http://t.co/cil5pvJS
I am curious-how would you describe all of the moves the Red Sox have made thus far?-Jerry
Today in Sports Plus: Lester impressing Farrell #constantcontact #RedSox http://t.co/d229d95d
#RedSox fans share your 2013 David Ortiz projections here- http://t.co/vAS43Ee6 #redsox -Jerry
Today in Sports Plus: Bill James sees bright future for #RedSox #constantcontact http://t.co/stydCQR1
Today in Sports Plus: Ross walks and #RedSox get - nothing #constantcontact http://t.co/oQOb3KIS
Today in Sports Plus: Competitive in 2013 #constantcontact #RedSox http://t.co/kGcEyvQt
Merry Christmas from Today in Sports #constantcontact http://t.co/TZVN3HEk
Merry Christmas Everybody!-Jerry
Today in Sports Plus: Slow news day for #RedSox #constantcontact http://t.co/r10UsqZ6
Today in Sports Plus: Farrell sees bullpen as "strong group" #constantcontact #RedSox http://t.co/fimWkxLw
Today in Sports Plus: Top 5 Red Sox stories from 2012 #constantcontact http://t.co/hbchWVKV
Today in Sports Plus: #RedSox - are they much better or just a littler? #constantcontact http://t.co/mxfFy5X0
Today in Sports Plus: #RedSox putting last year behind them #constantcontact http://t.co/X8srf3vw
Have fun tonight on New Year`s Eve and be safe-Jerry

SCREENNAME: Jerry_Remy

Sports & Recreation 28.047657
Baseball 18.833333
Arts & Entertainment 2.27015
Holidays & Celebrations 2.178451
Society & Culture 2.168723

Figure4.1 : User classification results for user with screenname Jerry_Remy with respective scores.

The user classification results for user *Jerry_Remy* who is a *Major League Baseball Broadcaster* shows that his expertise is *sports* , especially *baseball*. Scores of *Sports & Recreation*, *Baseball* are far more than scores of other categories, which shows that the user tweets only about *sports* and *baseball* most of the time,

There are numerous applications of user classification which are discussed later. In the project, we implement 1 application just to show the motivation and relevance of the problem solved. In the next subsection, we discuss this application of suggesting news articles to users.

4.1 News Recommendation System

This is one of the applications of user classification. Here, we plan to recommend news articles to Twitter user based on the classification of his tweets. These news articles would be recent news articles and would be pertaining to his interests which we have extracted using the proposed model.

Since the model uses only 100 tweets for user classification, it can be stated that the mentioned interests are his recent interests. *This makes our news recommendation system dynamic*. So, if a user who was earlier tweeting only about soccer because the soccer world cup was going on and is now interested in politics, economy, then his change of interests would be captured by our user classification model.

For suggesting news articles, we use Yahoo's BOSS Search API. BOSS is Yahoo!'s open search and data services platform. After obtaining classification of a user, we take the names of the yct categories in the user classification and query BOSS API for the recent news articles concerning these topics such as *Sports & Recreation, Politics & Government, Technology, Arts & Entertainment* . A web based application has been developed during the course of the project which takes the twitter handle of users as input and gives the user categories as output along with the recent news articles. Snapshots of the web application have been included in the section as Figure 4.2 and Figure 4.3 . These news articles could also be suggested based on the country of twitter user.

There are various other applications which are discussed in section .

Figure 4.2: Classification of Barack Obama and news recommendations

Figure 4.3: Classification of Tech Crunch and news recommendations

Chapter 5

Results and Analysis

In this section, the proposed model is tested and evaluated for various possible scenarios which are discussed in detail.

Pearanalytics which is an internet marketing tools and services company conducted a Twitter study in 2009 in search of finding out what people are really using Twitter for. The following results were obtained. Figure 5.1 shows categories with their percentage tweets in Peeranalytics study.

Categories	***Percent(%) of total tweets***
Total News	3.6
Total Spam	3.75
Self Promotion	5.85
Pointless babble	40
Conversational	37
Pass Alongs	8.70

Figure 5.1: % of tweets belonging to every category specified

The above results obtained by the survey shows that a majority of tweets are pointless babble and personal conversations. So, a very high recall can never be obtained because a significant amount of tweets of an average twitter user might be such that they cannot be classified. For eg: tweets like '*Good Morning !!! Have a nice day '* can never be classified into the subject categories which we have decided.

5.1 Model Categorization V/S Human Categorization

An interesting survey has been done as part of the project where we compare the proposed model against human categorization. Human categorization means we select a few random tweets and ask people to classify the tweets as having some topic or not having any topic. People can classify the tweets into any category they wish as right for the tweet. There is no complete list of categories available to those who take survey. But, the model classifies the tweets into Yahoo Content Taxonomy (YCT) categories.

A web page was put up where people were asked to take the survey where they had to classify 10 tweets as topical or non-topical to complete the survey. The survey was done on 400 tweets i.e. 400 tweets were classified by users as topical or non-topical. Then the same set of tweets were run against the model and results were compared. The results are as follows:

1. 180 tweets were classified by survey as topical.
   - Out of these, 43 tweets were classified as topical by the survey but didn't get any classification through the model.
   - The remaining 137 tweets were classified by both the model and survey.
   - Among the tweets classified by both the model and the survey, there were no misclassification in the survey but 23 tweets were misclassified by the model.

2. 234 tweets were classified as non-topical by the survey.
   - Out of these, 80 tweets have been classified into some categories by the model.

80 tweets which were not classified into any categories by the model were given back to some survey- takers for re-classification but this time they were provided with the list of entire yct categories and they classified 63 of these tweets into some categories.

The tweets for which both the classifications – the survey classification and the model classification are available have the same classification for *70 % of these tweets.*

Looking at figure 5.2 shown below, 217 tweets out of 400 got classification thorugh the model. This is approximately 50% of the tweets. The results of peeranalytics study show that a majority of a set of tweets is pointless babble and so, 50% classification is better.

		Classified by survey	
		Yes	**No**
Classified by model	**Yes**	137	80
	No	43	154

Figure5.2: Comparison of survey and model statistics

5.2 Number of Wikipedia entities

The model proposed here shows improvement over using CAP search directly using Wikipedia as an external knowledge base. We collected around 15 million tweets belonging to a group of users were analyzed for the month of June, 2009. Approximately, 12000 users posted more than 100 tweets. So, the total tweets under consideration are around 2 million. Here, we run all the tweets through CAP and find the number of Wikipedia entities present in each response. The significant number of tweets with Wikipedia entities show that the proposed model would definitely give better results as compared to using direct yct results from CAP.

The users who posted the tweets are chosen at random i.e., the dataset does not contain the tweets of only eminent personalities. It is a fair assumption that celebrities, famous people tweet sensibly as many people follow them. So, taking not only their tweets, but also the tweets of people picked at random, gives us a general and fair idea from the results obtained.

The next figure, Figure 5.3 shows the percentage of tweets in the month June, 2009 having no Wikipedia named entity, 1 named entity, 2 named entities etc.

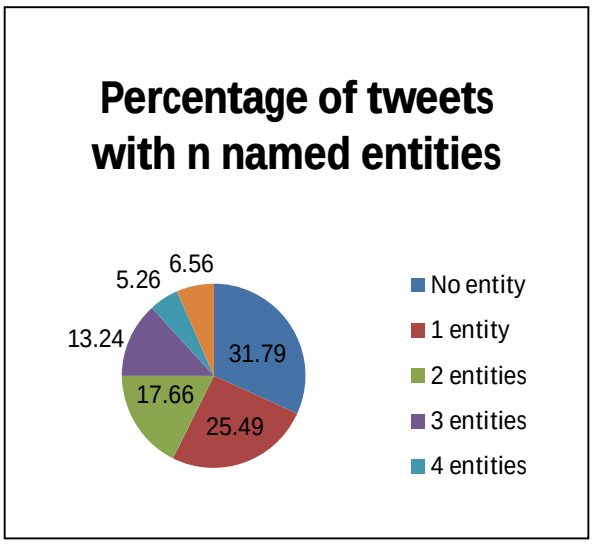

Figure 5.3: Comparison of number of Wikipedia entities

Tweets with no entity are considered to be messages which cannot be classified into some categories using Wikipedia. Some of these tweets could be classified directly to yct categories using CAP search. For instance, a tweet like '*hello everyone !!!!! have a nice evening*' doesn't have any Wikipedia named entity and can also not be classified into any topic category in general.

5.3 Wikipedia Classification v/s Yahoo Content Taxonomy Classification v/s URL

The approach that we discussed earlier uses direct yct categories obtained from CAP search and if Wikipedia entities are present, the model maps these entities to yct. To get an idea of how much improvement takes place when Wikipedia article mapping scheme is used over direct yct results. Direct yct means that yct categories are present in the response field of the content analyzer response field of the tweet. We also compare the URL results i.e. the cases when the tweets were not directly classified, not classified through Wikipedia but through classification of the URL present in the tweet.

We take a random sample of 1000 tweets for classification. These tweets are not restricted to any particular set of users, timeline etc. The tweets taken are in English language as we are classifying only English language tweets throughout the project. The tweets are fed

to Yahoo Content Analyzer (CAP) for direct yct categories. If yct categories are present in the response field for the tweet, such tweets are not examined further. If yct categories are not present and Wikipedia entities are present, those Wikipedia articles are mapped to yct categories if possible. Similar arguments go for the URL classification.

Here, we compare how many of 1000 tweets got classified without taking Wikipedia entities into account i.e. how many tweets got directly mapped via CAP and how many tweets needed Wikipedia knowledge for their classification.

In a sample of 100 tweets, 25 tweets got directly mapped to yct while 15 tweets got mapped through Wikipedia entities and 16 through URL classification. For a sample of 1000 tweets, 270 tweets got directly mapped to yct while 180 through Wikipedia entities and 63 through URL classification. So, a total of 503 tweets got mapped out of 1000 to yct. Also, here we are considering direct yct, wiki, URL sets as disjoint i.e. we search for wiki entity classification only of there is no direct yct classification and we search for URL classification only if there is no direct classification and no Wikipedia classification. So, it can be stated that out of the 270 directly mapped tweets, there might be a majority of tweets which have some wiki entity present in their CAP response fields.

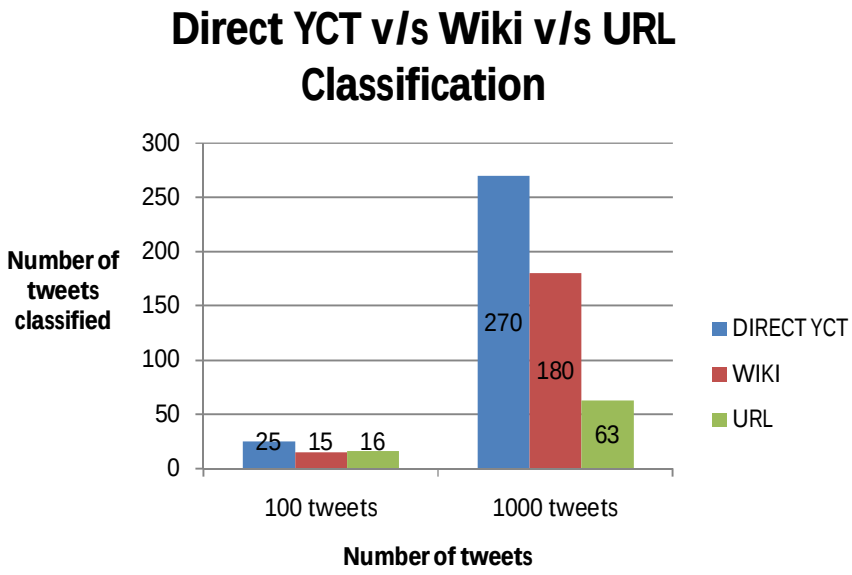

Figure 5.4 Comparison of Direct YCT v/s Wiki v/s URL classification

The above results in Figure5.4 show that when Wikipedia classification technique is used, then there is an improvement of around 20% over the model where we only use direct yct categories from CAP response field. Also, if both direct mapping and Wikipedia mapping give same results i.e. same classification categories, it can be used as a verification technique for correct classification.

5.4 Wikipedia misclassification

This is the last analysis which is done using classification of random tweets. After this, all the analysis involves classification of users and experts. We have introduced in section 4.3 how Wikipedia helps us in classifying more tweets than a direct approach. Here, we discuss about misclassification in Wikipedia article to yct mapping.

We have taken a sample of 50,000 Wikipedia articles related to various people, events, sports etc. at random. These articles are then mapped to yct categories. This should be kept in mind that this analysis is separate and doesn't take any tweet as input.

Out of 50k, no mapping was generated for around 6k articles which is around 14 percent no classification. For finding misclassification, mappings of 500 articles were manually verified. 20 articles out of 500 (4%) got misclassified. Here, misclassification means that they got classified into certain category which cannot be considered as the most important or representative thing regarding that article. For instance, in case of *Barack Obama*, he is also a *Grammy Award Winner* but this cannot be considered as his most important achievement.

5.5 Expert classification

In sections 5.1-.5.3, we have discussed results of tweet classification. In this section, user classification based on his tweet classification is discussed.

A random sample of 500 experts is taken along with their 100 tweets in taken. Most of these twitter users have at least one expertise i.e. they frequently talk about a specific interest such as *politics, economy, celebrities, entertainment, sports* etc. Most of these experts are not celebrities. Figure 5.5 shows the Twitter profile of a user in JSON format.

[0, 1152653177, 0, 0, "Publisher of Mighty Girl, co-founder Mighty Events. http://about.me/maggie", 978, 25243, 841, 1, 448, 0, "en", 1063, "San Francisco", "Maggie Mason", 0, "Maggie", 3874, "Pacific Time (US & Canada)", "http://www.mightygirl.com", -28800, 0, null, null, 1348453117, 250056615884120064]

Figure5.5: JSON tuple of a twitter profile

The fields of the tuple in Figure 5.5 are as follows:

contributors_enabled integer, created_at integer, default_profile integer, default_profile_image integer, description text, favourites_count integer, followers_count integer, friends_count integer, geo_enabled integer, id integer, is_translator integer, lang text, listed_count integer, location text, name text, protected integer, screen_name text, statuses_count integer, time_zone text, url text, utc_offset integer, verified integer, withheld_in_countries text, withheld_scope text, status_created_at integer, status_id integer

For the set of 500 experts that we take, we analyze their tweets during the month of *December 2012*. Every expert's tweets are passed through the proposed model for classification. After this, we use the tweeple classification technique discussed in Chapter 4. In this way, all 500 experts are classified into certain yct categories with some associated scores.

Now, these yct categories obtained for all the expert users are to be verified. For all these experts, words in their tweets are taken and all the stop words are removed. We call the remaining words in the tweets as *tokens*. Token list is expanded with every incoming tweet and number of occurrences of tokens are added with the tokens. Finally, we have a set of tokens along with their number of occurrences for a expert. We take the top 10 tokens i.e. 10 tokens with highest number of occurrences.

For verification, we compare the top 10 yct categories and top 10 tokens and try to find if there is any match between a category and a token. We say that an expert is mapped correctly if there are more than 2 matches with the token set.

Out of 500 experts, yct mappings were obtained for 474 users. Rest of the users didn't have sufficient tweets to be classified into yahoo content taxonomy. We match these 474 experts' yct results with their token sets discussed.

444 experts have same interests suggested in both yct mappings and their token set. 19 experts didn't have same interests suggested through their yct categories and their token set.

Figure 5.6 is a pie chart showing the percentage of experts classified correctly and misclassification. Figure 5.7 shows yct mappings and token set of 4 experts classified correctly.

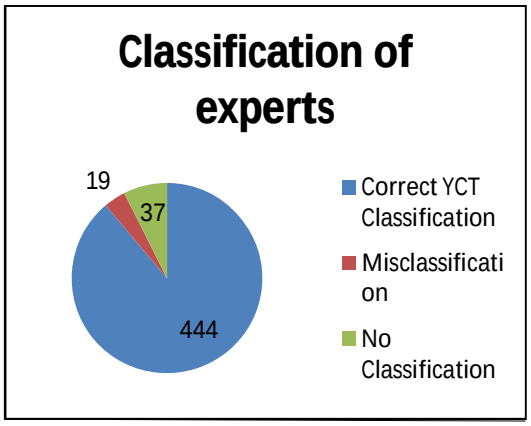

Figure 5.6

YCT MAPPING	*TOKEN SET*
SCREENNAME: holeinone *Golf 50.889952* *Sports & Recreation 46.11779* *Politics & Government 4.716528* *Celebrities 3.569505* *Arts & Entertainment 3.274227* *Holidays & Celebrations 2.333568* *Books & Publishing 2.043102* *Tennis 1.885728* *Travel & Tourism 1.649124*	*SCREENNAME: holeinone* *golf 14* *golfers 4* *news 2* *sports 2* *boys 1* *planet 1* *world 1* *game 1* *players 1* *grange 1*
SCREENNAME: LEETERRYNE *Politics & Government 13.003767* *Nature & Environment 4.024979* *Government 2.378201* *Society & Culture 2.048551* *Arts & Entertainment 1.979189* *Holidays & Celebrations 1.962962* *Debt 1.263158* *Natural Phenomena 1.182608* *Weather 1.148514*	*SCREENNAME: LEETERRYNE* *politics 132* *congress 120* *house 107* *republicans 58* *government 30* *news 24* *gop 24* *officials 23* *reps 20* *nebraska 19*
SCREENNAME: earthtimes *Environment 4.519995* *Nature & Environment 3.808751* *Arts & Entertainment 3.595051* *Living Nature 3.155787* *Politics & Government 2.649964* *Finance 1.944664* *Investment & Company Information 1.716128* *Government Agencies 1.4* *Employment & Career 1.161431*	*SCREENNAME: earthtimes* *news 112* *green 70* *environment 69* *media 37* *eco 28* *earth 23* *world 22* *sustainability 19* *science 17* *blogs 14*
SCREENNAME: 24thminute *Sports & Recreation 17.27942* *Soccer 8.501195* *American Football 4.457925* *Politics & Government 3.658608* *Arts & Entertainment 1.502045* *Baseball 0.969231* *Beauty 0.718182* *Society & Culture 0.709544* *Society & Culture Organizations 0.597884*	*SCREENNAME: 24thminute* *soccer 103* *football 36* *sports 30* *footy 25* *bloggers 23* *tfc 22* *toronto 21* *fc 15* *toronto fc 14* *news 13*

Figure 5.7: YCT Mapping and top tokens for four Twitter experts

5.6 Expertise Verification *(Expertise Given)*

We take different set of users with their expertise already mentioned. So, in this part we check whether we get the same expertise through the proposed model. We take users with expertise in *baseball, beer, comedy, economics, education, environment, golf, government, history, iphone, law and soccer*. 50 experts in each field are taken and then their expertise is verified through the model.

Figure 5.8 shows percentage of experts correctly classified.

Expertise	*Expert classified correctly(%)*
Baseball	*100*
Golf	*100*
Soccer	*100*
Beer	*68*
Books	*84*
Comedy	*59*
Economics	*100*
Government	*95*

Figure *5.8*

Most of the experts related to beer got classified into *Society & Culture, Dining & Nightlife*. The accuracy is low for *comedy* because most of the tweets of such users get classified into Arts & Entertainment as there is no dedicated yct category on comedy. Similarly, majority of experts related to comedy were mapped into *Arts & Entertainment, Sports & Recreation.*

This technique could also be used to verify dynamically if an expert is tweeting about his expertise or some other topics. Change in a user's expertise or interest can also be found using the proposed model.

5.7 Network of a Twitter Expert

In the project, we have not yet discussed the relationship between twitter users i.e. *follower* and *friend* relationship. If A follows B, then A is a follower of B and B is a friend of A. The twitter relationship graph is a directed graph i.e. the relationship is not both ways. In this analysis section, we take a set of twitter users. With every user, we associate some of his *friends* i.e. the people he follows. Now, we try to statistically support an argument that

A user is highly likely to follow people with whom he shares a common expertise.

The argument is supported by a set of 20 Twitter users. Every user in this set and his 40 friends are passed through the proposed model and categorized into yct categories. After this, we check for any similarity in their expertise manually.

Out of the set of 20 users, 8 users with screennames *slashfilm, SaraBareilles, vinotravelr, thot* seems to be following people with whom they share some common interest such as *sports, arts & recreation, software* etc. This suggests that if clusters of user interests are formed in twitter network are formed, then there is higher probability of an edge to lie inside a cluster as compared to edges between clusters.

In this chapter, we have supported the model's accuracy, importance and relevance through various techniques and statistical information. There are no comparison results with other models as some of the other techniques do not employ any external knowledge base and so fare poorly. There have been almost no technique to my knowledge where a predefined elaborate set of user interests is taken as classification categories.

The next chapter discusses some of the applications of our proposed model.

Chapter 6

Applications of Proposed Model

The model that has been discussed classifies tweets into certain yct categories and then classifies users into these categories based on his tweets. User classification has various applications which we discuss in this chapter. We have already discussed one of such applications – **News Recommendation System** in Chapter 4. This application has been implemented through a web interface. Now, we discuss some of the applications which could use the proposed model heavily.

1. **News Recommendation System –** This application has been developed as part of the project. This recommendation system is discussed in detail in section 4.1. Recent news articles with their links are recommended to Twitter users based on the model classification.

2. **Targeted Marketing** - A target market is a group of customers that the business has decided to aim its marketing efforts and ultimately its merchandise towards. Twitter has a lot of information regarding a user stored in his tweets. If the expertise of a user or his interests are somehow obtained, then those interested could be used in targeted marketing. For example, if a company launches a new *electronic gadget*, then it can circulate it to all those Twitter users who have an expertise in *electronics*. So, this can be used in advertising, suggesting gifts etc. Decent and specific marketing can be done and spamming can be avoided.

3. **Interests based Twitterfeed –** Currently, tweets are chronologically ordered in a user's timeline. Using user interests found by the model, any incoming tweet can be classified into certain categories and if these categories match user interests, then such a tweet could be given priority in the tweet feed. This would make browsing tweet stream for interesting content a lot easier for various users.

4. **Suggest friends –** Twitter users who are looking to follow other users with similar interests can benefit from such a model. If users are classified into certain categories, then if a user wishes to follow some new people, users can be recommended to him based on his user interests. In this way, user can easily find and follow people who share the same interests.

5. **Suggest experts –** Tweeples who are looking to follow experts related to various specific topics such as *soccer, golf, software, politics* etc. will find the proposed classification technique helpful. If a user wishes to follow people who are interested in various cuisines can follow experts in the category *Cooking, Dining & Nightlife* etc.

6. **Tweet specific sending –** Currently, when a user posts a tweet, that tweet goes to all his followers. Using this feature, an additional optional can be made available to send a tweet only to the followers who share the same interests as the categories of the particular tweet. For eg: A tweet related to last night's soccer game can be sent only to followers interested in soccer.

The above mentioned applications are some broad areas where the technique could be used. User classification and tweet classification is a much needed functionality in Twitter data because most of the data is unstructured but has huge information regarding a person which could be used intelligently for user's own purposes.

Chapter 7

Conclusion

The thesis aims at classification of a tweet into certain interest categories. This is later used in proposing a technique for user classification based on classification of his tweets. For the classification, Yahoo Content Taxonomy(YCT) categories are used and Yahoo API is used for a direct classification if possible. Most of the tweets are not directly mapped to yct categories which led to incorporation of a knowledge base- Wikipedia. Wikipedia article information and Wikipedia taxonomy is then heavily used for tweet classification. The accuracy and relevance of the model is supported by various statistical experimental results. The model has a pre-defined set of around1000 yct interest categories which are necessary as well as sufficient. The approach proposed is perhaps the only approach which uses Yahoo Contextual Analysis Platform for tweet classification with improvisation done using Wikipedia as an external knowledge corpus.

REFERENCES

1. Phan, X. H., Nguyen, L. M., and Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. Proceeding of the 17th international conference on World Wide Web. pp. 91- 100. ACM, (2008).
2. Go, A., Bhayani, R., and Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford. (2009).
3. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M.: Short text classification in twitter to improve information filtering. Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval. pp. 841-842. ACM, (2010).
4. Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A.: Twitter power: Tweets as electronic word of mouth. Journal of the American society for information science and technology. 60, 11, pp. 2169-2188. Wiley Online Library, (2009).
5. Stone, B., Dennis, S., and Kwantes, P. J.: Comparing Methods for Single Paragraph Similarity Analysis. Topics in Cognitive Science. Wiley Online Library, (2010).
6. Banerjee, S., Ramanthan, K., and Gupta, A. Clustering short text using Wikipedia. In Proc. SIGIR (Amsterdam, The Netherlands, July 2007), 787-788.
7. Hu, X., Sun, N., Zhang, C., and Chua, T.-S. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In Proc. CIKM (Hong Kong, China, Nov. 2009), 919-928.
8. Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, and M. D., Sperling, J. TwitterStand: news in tweets. In Proc. ACM GIS'09 (Seattle, Washington, Nov. 2009), 42-51.
9. Michelson, M. and Macskassy, S. A.: Discovering users' topics of interest on twitter: A first look. Proceedings of the Workshop on Analytics for Noisy, Unstructured Text Data. (2010).
10. Liu, X., Zhang, S., Wei, F., Zhou, M. : Recognizing named entities in Tweets. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 359–367, Portland, Oregon, June 19-24, 2011.
11. Benson, E., Haghighi A. and Barzilay R: Event discovery in social media feeds. In ACL '11, 2011.
12. Genc, Y., Sakamato, Y. and Nickerson, J: Discovering Context: Classifying tweets through a semantic transform based on Wikipedia
13. Meij, D., Trieschnigg, M. de Rijke. and Kraaij, W: Conceptual language models for domain-specific retrieval. Inf. Process. Manage., 46(4):448–469, 2010.